# Twitter Recommender Systems: A topological and interest oriented approach

**ABSTRACT:**

Twitter has become a big buzz word nowadays. The plethora of tweets published every day gives a big data to work on. Much research has been done in this field of recommendation. In this paper we propose a recommendation system where the dynamic level of interest of the user is used as a factor and the potential users selected for recommendation to a user is based on the topological relation between that of the user and the recommended one. Such recommended tweets are found to be more appealing to the user. The proposed system targets the user's interests, which change dynamically over the time, and recommend the tweets that correspond to such dynamic interests.

**Keywords:** Twitter Tweets, Dynamic Level of Interest, recommendation system

## 1. INTRODUCTION:

Twitter is an online social networking service that enables users to send and read short 140 character messages called "tweets". It was created in March 2006 and launched by July 2006. Registered users can read and post tweets. The twitter service gained worldwide popularity with 100 million users by 2012 who posted 340 million tweets per day. In 2013 twitter was one of the ten most visited websites and was described as "SMS of the internet". As of 2015 Twitter has 302 million active users. [1]

Twitter allows its users to follow other user's tweets. User's retweet, favorite, reply to tweets that are interesting to them. Users have to go through a lot of tweets to find that which are relevant to them. Recommending the right tweet which is relevant to the user's interest is a great challenge in itself.

Recommendation systems have helped users with providing interested tweets to them. Our aim is to recommend tweets to users based on formation of a topological network of followees-followees, further optimizing the network to find similar users, dynamic level of interest the user has on different topics he/she interested in and finally optimize the recommendation based on implicit, explicit and additional features of the tweet. In this approach we consider the assumption that a user's interest is dynamic.

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 is about the problem description, Section 4 describes about block diagram. Section 5 describes about methodology. Section 6 describes about the implementation. Section 7 concludes the paper and shows direction for the future work.

## 2. RELATED WORK

As our approach focus on tweet recommendation and topic modeling, we provide an overview of the related research. To overcome the information overload problem in online social networks, many recommender systems were introduced to help users find interesting information. Some of these systems were conducted to study the social network structure and recommend friends to the user based on the similarities of interests. One study proposed Twittomender that recommends Twitter users based on the relationships of their Twitter social graphs [4]. Kwak estimated the influence of users on Twitter by proposing three methods: the number of followers, Page Rank and the number of retweets [8]. Other systems studied the personalized recommendation systems to recommend only useful content to users. Chen et al. proposed a collaborative filtering method to generate personalized recommendations in Twitter through a collaborative ranking procedure [3]. Pennacchiotti et al. proposed a method to recommend tweets to users by following users' interests and using the tweet content [7]. Chen et al. proposed a URL recommendation model to demonstrate the utility of various combinations of tweet content and social graph information during recommendation [6]. Abel et al. analyzed how user's profiles changes over time, and how to recommend news articles for topic based profiles. Our model is different in that it tries to capture the change of each user's interest in different topics over time, and recommend interesting tweets based on this interest [5]. The study done by Shaymaa Khateret al is most relevant to our problem finding dynamic level of interest is discussed in [1]. Forming a topology for a given user for followee recommendation and recommending tweets based on them is discussed in [2].

## 3. PROBLEM DESCRIPTION:

When users login on Twitter, they see a stream of tweets sent by friends which composes their timeline. Many of these tweets are conversational tweets and/or are not of personal interest to the user. The goal of our model is to decide for each user the tweets that might be of interest from the user's timeline. Besides being able to post their own tweets, users can also interact with their timeline by replying, retweeting or favoring the tweets. As there are no explicit means to extract the user's level of interest in a tweet from Twitter, we relied on these actions to predict the user's interests. Hence, the retweets, replies and favorites can be used as an indication of the interest of the user in the corresponding tweets.

## 4. BLOCK DIAGRAM:

Proposed work consists of dynamic level of interest from topologically derived users. We propose a system where users are suggested based on topology and also tweets among the users are recommended based on dynamic level of interest. In addition, Feature based extraction in which tweets are also recommended based on features like user publisher similarity, role of favorites and retweets, etc is also considered.

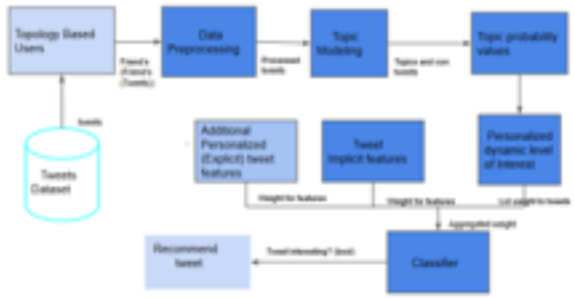

**Fig 1. Block diagram**

## 5. DATA COLLECTION:

The data is collected using twitter's REST API. This API has a restriction of 15 requests per 15 minutes. 3 base users where selected from diverse fields whose most relevant list of friends are extracted. From the extracted list the friends of them who are close to the base user is selected. The features of the tweet extracted and features of the users extracted are mentioned in table below.

**Table 1. Features of tweets and users**

| Tweet Features | User Features |
|---|---|
| Created at | User name |
| Tweet Text | User id |
| Time of Publishing | Status |
| Tweet id | Number of tweets |
| Hashtags | Account Age |
| Number of Favorites | Friends count |
| URL | Followers count |
| Length of Tweet | Favorites count |

## 5. METHODOLOGY

### 5.1. TOPOLOGICAL FORMATION:

To recommend users to a given base user. A topological graph should be formed to extract the friends who interest her and friends of those interesting friends who have similar interests to
The user to be recommended.

### 5.1. TOPIC MODELLING:

As stated in [1] we can predict user's interests by classifying them into topics. According to the frequencies of words appearing in each document the model determines the relevant set of words to each topic. As described in the algorithm 1 in [1] a set of posts is classified as $E = \{ e_1, e_2, ..... e_n \}$, the modeling tool generates a set of topics denoted by $L = \{ l_1, l_2, .... . l_k \}$

### 5.2. DYNAMIC LEVEL OF INTEREST:

In this section, we study how the interests of individual users about a certain topic change over time. Getting the dynamic level of interest in a tweet takes place through some steps:

1. First we get the per topic activity in each day d for the user, denoted by $A_d = \{ a_d^1, . . . , a_d^k \}$ where $a_d^i$ is the level of activity of the user in topic $l_i$ on day d. $A_d$

is calculated by adding the vectors o e in that day, as in Equation 1. The details of this step are shown in Algorithm 1 in [1].

$$\mathbf{A}_d[i] = a_i^d = \sum_{\forall e \in E: e_{date} = d} \mathbf{o}^e[i] = \sum_{\forall e \in E: e_{date} = d} o_i^e \quad (1)$$

2. Given a new tweet, the user's level of interest in the tweet can be calculated using Equation 2. Basically, the equation sums up the user activity vectors in the window of last seven activity instances prior to the tweet creation day d. Each of these instances corresponds to user's actions done in one day. For a user who is active (posting a tweet, replying, retweeting or favoring another tweet) every day, this window will span one week period. For less active users (not active every day), this window will be longer to cover the last seven active days in which the user was active. We only consider the last seven instances, as considering intervals longer than seven days will introduce irrelevant noisy tweets. This step is illustrated in Algorithm 2[1]

$$\begin{aligned} LevelofInterest(u, e_{new}) &= \sum_{l \in L} (\mathbf{o}^{e_{new}}[l] \cdot \sum_{d \to d-7} \mathbf{A}_d[l]) \\ &= \sum_{l \in L} (o_l^{e_{new}} \cdot \sum_{d \to d-7} a_l^d) \quad (2) \end{aligned}$$

```
Algorithm 2
function CalculateLOI(User u, Tweet e)
begin
   $o^e$ ← Percentages of topics in e from LDA model
   d ← e posting date
   LoI ← 0
   for each Topic l in L do
            val ← 0
            for i = 1 to 7 do
               val ← val + $A^u_{d-i}$ [l]
            end for
            val ← val * $o^e$[l]
            LoI ← LoI + val
   end for
return LoI end
```

## 5.3 PERSONALISED TWEET RECOMMENDER

In addition to measuring the dynamic level of interest for each user, some other static features can affect his interests. Some of these features represent the personalized interests of the user, others are general features that are related to the tweet‘s quality or the publisher‘s authority that can affectthe tweet‘s degree of interest to the user. The following sections describe the personalized features and other explicit features that might affect user's interest.

### 5.3.1 Personalized Social Features

Social features are the features that represent the social relationship between the user and the publisher. This relation can be friendship, neighborhood who posts tweets about events happening in the neighborhood or celebrities who have interests in common with the user. These social features include

• **User-publisher similarity:** this feature measures the similarity between activity level of the user and the publisher on all topics. This is measured as the cosine similarity between vectors formed by summation of the level of interest in a topic for the user over time. This is shown in Equation 3. Generally, the cosine similarity measure yields a value between -1 and 1. The value of 1 means the exact distribution match, i.e., activities of both users are distributed in the same proportions on different topics, though one of them might be generally more active than the other. The value of 0 means that users have nothing in common.

### 5.3.2 Explicit Features

Besides the implicit features, we also looked at other explicit features such as publisher based , tweet based features. This information was got from Twitter API using various API calls.

• **Publisher based features:** This refers to the tweets owner's features such as publisher followers, publisher tweets count, mention count.

– **Publisher followers:**We took the count of the follower the publisher has and assigned various weights to it. When the publisher has many followers it means his tweets are highly recommendable . Hence higher the number of followers, greater the weight assigned to it.

– **Publisher tweets count**: We took the number of tweets posted by the user from the beginning of his account's existence. When the user has tweeted many tweets it means he uses Twitter actively. Hence larger the number of tweets, greater the weight assigned to it

– **Mention count:** We also took the number of times the user is mentioned in entire tweet base. This reflects the interaction level and popularity of the user. Thus higher weight is assigned to user having higher mention count.

**• Tweet based features:**

Tweet based features like retweet count, hasURL, hasHashtag were also considered.

**– Retweet count:** the number of times a tweet is retweeted is considered here. This reflects the popularity of the tweet among people. Hence larger the number of tweets, higher the weight assigned to it.

**– HasURL, HasHashtag:** sometimes a publisher includes supplement to their tweets with URL or hashtags. Hashtags can sometimes be an indication of the tweet's topic. The number of hash tags and urls included in the tweet is also considered. We assume that a tweet which has higher number of hashtags or urls is more informative than a tweet which has lesser number or none at all.

### 5.3.3 Additional personalized tweet features

Additional personalized tweet features such as user relationship feature, content relevance feature, content based, publisher authority features were also included.

**User Relationship Features:**

These features refer to the relationship between the publisher and the base user.

• Co-Friends Score: This score is dependent on the number of friends the publisher and the base user has in common.

• Co-Follow Score: This score is dependent on the number of followers the publisher and the base user has in common. Higher the common followers, greater the score.

• Mention Score: This score is dependent on the number of times the base user has mentioned the publisher in all his tweets. This reflects the like-mindedness, common interest of the base user and publisher. Higher the number of times, larger the score.

• Retweet Score: This score is dependent on the number of times the base user has retweeted the publisher in all his tweets. This reflects the interest of base user in the publisher's topic of interest. Higher the number of times, larger the score.

• Mutual Friend: This score which is either a 0 or 1 is dependent on whether the publisher and base user are friends or not.

**Content-relevance Features**:

Various content relevance features like neighborhood relevance, retweet relevance, relevance to hash tags and URL Domain history.

• Content Relevance: R(t, CP(u)), reflects the relevance of the plausible recommendable tweet to the user's content profile. Where CP refers to the content profile of the user and t refers to the incoming tweet.

• Neighborhood Relevance: R(t, NP(u)), reflects the relevance of the plausible recommendable tweet to the user's neighborhood profile, where t refers to the incoming tweet and NP refers to the neighborhood profile.

• Retweet Relevance: R(t, RP(u)), measuring the relevance of the plausible recommendable to the user's retweet profile, where t refers to the incoming tweet and RP refers to the retweet profile. This feature estimates the relevance between a tweet T and the retweeted history of a user U. Similarity scores between T and every post retweeted by U are calculated and summed to get the relevance score

- Relevance to Hash Tags: R(t, HT(u)),reflects the relevance of incoming tweet to the user's history of used hash tags, where t refers to the incoming tweet and HT refers to the hash tag history of the user.

- URL Domain History: R(t,UH(u)), reflects the relevance of incoming tweet to the user's history of URL's, where t refers to the incoming tweet and UH refers to the URL history of the user.

**Content-based Feature**:

Content based features like length of tweet, hash tag count, hash tag history, url count.

- Length of Tweet: The length of tweet is considered here. When the tweet is longer, we assume that it is recommendable because the user has spent more amount of time to write the tweet.

- Hash Tag Count: The number of hashtags is considered here. When the tweet has many hash tags it means, the tweet is more recommendable because the user has spent more time to include tags in the tweet.

- Hash Tag History: How many times the hash tag appears in user's retweets.

- URL Count: The number of URLs used in tweet i. On Twitter, publishers often include a URL as a supplement in their tweets. This is because the publisher cannot summarize their information in 140 characters and they refer URLs for fuller information on another web site. The number of URLs in a tweet is estimated by this feature

**Publishers' Authority Feature**:

These features potentially characterize how popular and how well connected a user is. Intuitively, a popular user who has many friends and followers can be actively passing information by retweeting messages

- Account Age: Number of years user u appeared on Twitter

- Friend Count: The number of friends user u follows. This feature records the number of people who follow the publisher. It is an objective measure of the popularity of a user based on public opinion.

## 6. IMPLEMENTATION:

In this section, we describe our datasets and the preprocessing steps followed by the experimental results for each step in our model.

### 6.1. TOPOLOGY BASED DATA COLLECTION:

Existing work on topology based recommendations is without modeling the interest of the user, thus bringing in a wider range of tweets which might be of dis-interest to the user. In our work, we topologically collected friends which are limited based on topic modelling.User's friends have very similar interest to the base user. The fact that his/her friends' friends might also have similar interest is implemented in this approach. We sourced our data from Twitter.com using REST api.Data of the following were collected: Base user's tweets and details, followee's details, followees'sfollowee's details and tweets. The topology for a given user is generated by creating friends of friend's network. A person's followees (friends) are chosen and the friends who best suite the interests of the given user are identified. Over the select users, their friends are selected. In the set of friends of friends the person which best suites the interest of the base user is identified. Their tweets are collected over a period of time preprocessed and the tweets which interest the base users are recommended.As Twitter APIs does not allow access to the timeline of the user directly without authorization, we build each user's timeline by first getting the posts for each of the base users, and then following the tweets posted by their friends, and consider them the scanned tweets by the user. All the favored tweets by the base users are also retrieved.

### 6.2 DATAPREPROCESSING

We preprocessed the tweets by discarding tweets with non English words. We also removed meaningless words such as stop-words, Twitter specific stop words, user names, and special characters and stemmed the remaining words. Data was preprocessed to segregate into the following files: Tweets alone, hashtags, mentions, retweets, description/status of user.

We build our model from a repository of more than five million tweets. To eliminate incomplete and noisy data,

### 6.3 TOPIC MODELLING:

Input of processed tweets of base user is given. Assumption is that input tweets are in a csv and contains the entire tweets of the users without hashtags and urls.Stop words are removed, suffix stripping, stemming is performed. Topics which interests or relevant to the user is output using the Latent Dirichlet allocation Tool. For topic extraction, Google's Topic modeling tool [8] was used to build topics from the given corpus.

### 6.4 CALCULATING DYNAMIC LEVEL OF INTEREST

In real life, the degrees of popularity of the topics are not constant. There are topics that attract more users than the others. Also, the user's interest in one topic can change from one time to another. The dynamic level of interest (LoI) is calculated using Equation 2. The user's dynamic LoI is based on the dynamic level of activity of the user in each topic. Input of tweets along with topics extracted using topic modeling is given in this module. Assumption is that user is active in the given period.Tweets from LDA o/p within particular time period chosen is aggregated with date information Output is thepercentage of topics in given tweets

### 6.5 INTEREST IN NEW TWEET

Input of topics with their probability values is given as input to the system. Assumption is that users have a dynamic level of interest that changes over a period of time. Input tweet is topic relevant to LDA o/p.Tweets from recommendable users are chosen based on cosine similarity of user profile. Level of activity in a topic and level of interest in a new tweet is the output from this module.

### 6.6 FEATURES

Input of tweets categorized as plausible for recommendation and a new tweet is given to this module. Booleanoutput for the new tweet i.e., either recommends it or not to the base user is configured. Algorithms likeRandom forest, Decision table, SVM, Linear regression were used to evaluate the performance of the recommender system.

### 6.6 RESULTS

The results of various models were compared and the accuracy was measured using parameters like Mean absolute error, Root mean square error, Relative absolute error, Root relative squared error. Mean absolute error (MAE) reveals how close forecasts or predictions are to the eventual outcomes[9]. (I.e. how correct the recommended tweets are to the actual interest of the user). The **root-mean-square deviation (RMSD)** or **root-mean-square error (RMSE)** is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. Low RMSE value for linear regression indicates that the values observed are very close to that of sample population. Root relative squared error (RRSE) gives the highest error value .The **relative absolute error** is very similar to the relative squared error in the sense that it is also relative to a simple predictor, which is just the average of the actual values. In this case, though, the error is just the total absolute error instead of the total squared error. Thus, the relative absolute error takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor. Support Vector machines (SVM) and linear regression gives the least error which indicates that these models are accurate in classification.

| Classifier model | Correlation coefficient | Mean absolute error | Root Mean Square error | Relative absolute error | Root relative squared error |
|---|---|---|---|---|---|
| Random forest | 0.9586 | 15.2 | 24.7 | 18.6 | 28.5 |

| | | | | | |
|---|---|---|---|---|---|
| Decision table | 0.9687 | 13.1 | 21.4 | 15.9 | 24.7 |
| SVM | 0.9703 | 12.7 | 21.3 | 15.4 | 24.6 |
| Linear regression | 0.97 | 13.3 | 21.0 | 16.2 | 24.2 |

**Table 2. Graphical comparison of models**

## 7. CONCLUSION AND FUTURE WORK:

As innumerable amount of junk tweets are produced every minute in twitter. Recommending users with the right kind of tweet which interests the users becomes necessary. We've implemented topologically relation based method to extract users which might interest the user and recommend the user the tweets which have similar interest to the base user. The concept of dynamic Level of Interest is used to recommend the best tweets. Results have concluded that on classifying there is an improvement in the average precision. The model analysis showed that the model has higher gain for users with high activity level.

As a future extension of this work, one can understand the mental behaviour of users using tweets, do a general analysis on favourites, retweets, hash-tag , or reply analysis, understand the network of a user, build social graphs, identify relationship between users as follower-followee/ retweeter/ favoriter, predict list of people who will favourite a given tweet, identify time spent by users on social network, analyse usage pattern, perform influential user detection, URL analysis, analyse composition of favourite user i.e. check if it is the same set of users favoriting every time / mutual favoriting takes place ,or answer hypothetical questions like if the time interval between tweets affects conversation length , or if active users on Twitter gains more favourites/retweets, predict trust links among users, identify retweet patterns of users among different usage pattern ( information seeker, social chatter etc.)

## 8. LIST OF REFERENCES:

[1] Shaymaa Khater , Hicham G. Elmongui , Denis Graˇcanin,Tweets You Like: Personalized Tweets Recommendation based on Dynamic Users Interests, *ASE BigData/SocialInformatics/PASSAT/BioMedCom 2014 Conference*, Harvard University, December 14-16, 2014

[2]Marcelo G. Armentano, Daniela Godoy and Analía Amandi,Towards a Followee Recommender System for Information Seeking Users in Twitter

[3] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative personalized tweet recommendation. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, pages 661–670, New York, 2012. ACM.

[4] J. Hannon, M. Bennett, and B. Smyth. Recommending Twitter users to follow using content and collaborative filtering approaches. In Proceedings of the fourth ACM conference on Recommender systems, RecSys '10, pages 199–206, New York, 2010. ACM. [8] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In Proceedings of the 19th International Conference on World Wide Web, WWW '10, pages 591–600, New York, 2010. ACM.

[5]F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on Twitter for personalized news recommendations. In Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization, UMAP'11, pages 1–12, Berlin, Heidelberg, 2011. Springer-Verlag.

[6]J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, pages 1185–1194, New York, 2010. ACM.

[7]M. Pennacchiotti, F. Silvestri, H. Vahabi, and R. Venturini. Making your interests follow you on twitter. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, pages 165–174, New York, 2012. ACM

[8] https://code.google.com/p/topic-modeling-tool/0

[8]Weka Tool http://www.cs.waikato.ac.nz/ml/weka/

Ancy Philip is doing her final year of bachelors in engineering at college of Engineering, Guindy. She has been a great learner and topper. She was awarded Ford's

Allan Mullaly Leadership in Engineering award for her brilliant pursuance in Leadership and Engineering.

Elaiya Bharathi is doing his final year of bachelors in engineering at College of Engineering, Guindy. He has been a great android developer and has won many hackathons throughout the college period. He has a great penchant for learning and has won many prizes in college competitions.